signals

MDPI

Article

Voice Transformation Using Two-Level Dynamic Warping and Neural Networks

Al-Waled Al-Dulaimi *, Todd K. Moon and Jacob H. Gunther

Department of Electrical and Computer Engineering, Utah State University (USU), 4120 Old Main Hill, Logan, UT 84322-4120, USA; todd.moon@usu.edu (T.K.M.); jake.gunther@usu.edu (J.H.G.)
* Correspondence: alwaled.aldulaimi@aggiemail.usu.edu

Abstract: Voice transformation, for example, from a male speaker to a female speaker, is achieved here using a two-level dynamic warping algorithm in conjunction with an artificial neural network. An outer warping process which temporally aligns blocks of speech (dynamic time warp, DTW) invokes an inner warping process, which spectrally aligns based on magnitude spectra (dynamic frequency warp, DFW). The mapping function produced by inner dynamic frequency warp is used to move spectral information from a source speaker to a target speaker. Artifacts arising from this amplitude spectral mapping are reduced by reconstructing phase information. Information obtained by this process is used to train an artificial neural network to produce spectral warping information based on spectral input data. The performance of the speech mapping compared using Mel-Cepstral Distortion (MCD) with previous voice transformation research, and it is shown to perform better than other methods, based on their reported MCD scores.

Keywords: voice transformation; dynamic warping; phase reconstruction; artificial neural networks

Citation: Al-Dulaimi, A.-W.; Moon, T.K.; Gunther, J.H. Voice Transformation Using Two-Level Dynamic Warping and Neural Networks. *Signals* **2021**, *2*, 456–474. https://doi.org/10.3390/signals2030028

Academic Editors: Tom Diethe and Niall Twomey

Received: 12 April 2021
Accepted: 13 July 2021
Published: 14 July 2021

1. Introduction

Voice transformation (VT) refers to the process of changing speech so that speech uttered by one speaker (the source speaker) sounds as if another speaker (the target speaker) had spoken it, for example, transforming from a male voice to a female voice [1]. VT has applications such as text-to-speech synthesis (TTS), international dubbing, health-care, multi-media, language education, music, security-related usage, vocal restoration, speech-to-speech translation, and preprocessing for speech recognition, etc. [2–4].

The general framework for VT is shown in Figure 1 [5]. In a training stage, performed offline, features are extracted from both source and target speech sources. VT, which maps information from the source speech to target speech, is trained to form conversion rules. In the actual operation, these conversion rules are used to convert features from a source to a target.

In this work, the spectral features computed using the DFT are transformed using a neural network (NN). Speech signals are aligned using dynamic warping (DW) which aligns both temporally and spectrally to produce training data which is used to train a NN. The NN is trained to produce the spectral mapping functions determined by the frequency-domain warping. Several NN structures (differing numbers of layers, and differing numbers of neurons per layer) were examined to determine which had best performance, as measured by mel-cepstral distortion (MCD). On the basis of MCD, this method was compared with previously developed methods, and is shown to perform well. The test set was based on transformation of a male speaker to a female speaker. Future work will test this architecture transformations between different speakers.

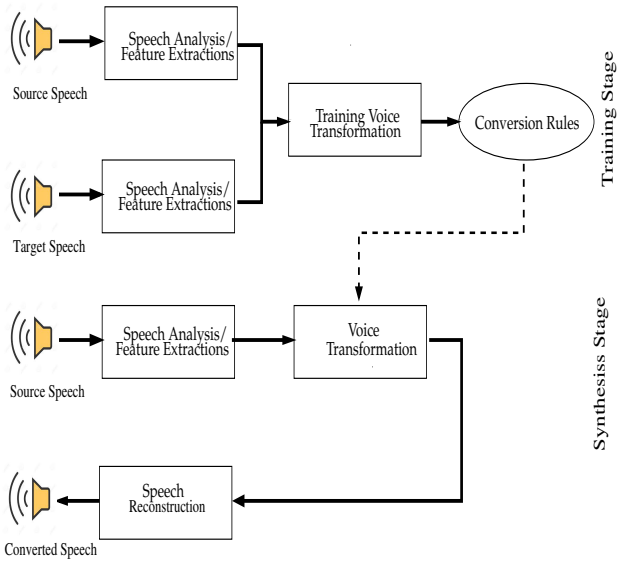


Figure 1. Typical voice transformation system.

2. Previous Voice Transformation Research

Early work on VT dates back to 1990, when the author in [6] proposed using a vector quantization approach, where the vector-quantized speech features are mapped spectrally via codebook mapping. The features transformed include acoustic features such as pitch, formant frequencies and bandwidth, spectral tilt, etc. [7,8]. These features are extracted and represented using vector quantized codebooks. Generally, codebooks are generated for both source speakers and target speaker. DTW is used to find the correspondence between these vectors. These correspondences are then accumulated to find a histogram, which is used as a weighting function to compute the mapping codebook from source speaker to a target speaker. In the synthesis step, the speech iwas quantized using the speaker's codebook, then all feature parameters are transformed with the mapping codebook from the training step, and finally the speech is synthesized with LPC vocoder.

To improve this general approach, the authors in [9] propose a method of doing VT based on Pitch Synchronous Overlap and Add (PSOLA) to improve the quality of the converted speech signal. In this work, the authors introduced DFW to achieve VT.

Gaussian Mixture Models (GMMs) have been used for VT [10–17]. GMMs model the source and target data with a GMM to produce transformation functions for each Gaussian component [11]. Two approaches have been used for GMM-based VT techniques: first modeling the source and target information separately using a GMM [18], and second modeling the joint density GMM (JDGMM) between source and target speakers [10]. The conversion function is typical assumed to be probabilistic piecewise linear mapping function for each Gaussian. The unknown parameters are calculated by solving the normal linear conversion equations for a least squares solution [18]. The combination of the spectral source aligned vectors and the corresponding target aligned spectral vectors is used to estimate GMM parameters for the joint density model by using expectation maximization algorithm [11]. Although this algorithm works effectively, its conversion performance is still insufficient [19]. The performance of the standard GMM based VT method have been improved by implementing sophisticated techniques, such as Gaussian process regression [20,21], also using WaveNet vocoder in GMM-based voice conversion to improve the naturalness of the synthetic voice [15,16]. An improvement can be achieved by incorporating a trajectory-based conversion algorithm to achieve temporal correlation in the conversion process [18]. Additional features were modeled in order to reduce the over-smoothing effect of the transformed signal parameters, such us transformation based on matrix variate GMMs, in which joint features from the source and target speakers are represented as matrices using multiple frame features [14], global variance [22] and

modulation spectrum [23]. Furthermore, a real-time conversion process has also been successfully implemented for state-of-the-art GMM-based VT [24].

As an alternative to the GMM, which relies on probabilistic mappings between the source and target parameters, DFW seeks to warp the source spectral information in frequency, in order to compensate for the differences between the spectral features of the source and the target speakers, to most closely match that of the target speaker [25]. Cepstral information was extracted from both source and target speakers to do DFW [26,27]. In [15] a new framework was proposed by combining frequency warping and exemplar-based method to achieve VT, where the target was represented by applying frequency warping on the high-resolution spectrum. The warping function is generated by a sparse interpolation from a dictionary of exemplar warping functions.

Neural networks have also been used for VT, where the NNs map the source speech feature vectors into the feature vectors of the target speaker [28,29]. The work [2] is an early effort using NN to transform the source speaker formants to target speaker formants using a feedforward NN. In [30], the authors proposed a method for VT using NN with three layers. This method is based on LPC spectral features using radial basis function NN. In [31] the authors compared using a NN and GMM for VT. GMMs capture the joint distribution of the source features and the target feature, while the work in [31] directly maps the spectral source feature information onto the spectral target feature information. Additionally, with GMM they use maximum likelihood parameters generation (MLPG) to obtain a smooth trajectory of spectral features, while the mapping with NN provide best transformation results without using MLPG.

Long short-term memory (LSTM) NNs have been used for VT [32]. The joint use of deep bidirectional (DB) LSTM and identify vectors (i-vectors) [33], Kullback–Leibler divergence and deep NN-based approach [34], and DBLSTM based Recurrent NNs [35,36], have been recently used for VT. More recently, Generative Adversarial Networks (GAN) such as VAW-GAN [37], CycleGAN [38,39], and many-to-many mapping with StarGAN [40] further advance the state-of-the-art.

Previous methods generally work by extracting a parametric acoustic parameter (such as pitch, MFCC, etc.) and transforming this parameter from one speaker to another based on estimated joint densities. This introduces complexities associated with this initial parameter estimation. The approach taken here avoids the need to find parametric acoustic parameters, or to model the joint density, and instead deals directly with spectral information. (That is, this is analogous to spectral estimation, where a parametric model might be used, such as AR or ARMA, or the spectrum may be estimated directly using the DFT.) NNs are used to learn the transformation which maps the spectral information of one speaker to another speaker. We have found that this basic idea yields improvements when the spectral information from the preceding and following speech segments are used to provide context. The transformation is achieved using a two-level DW. Based on the two-level DW it is straightforward to map the source speech to target speech when both are available.

3. Voice Transformation Process

3.1. Two-Level Dynamic Warping

The transformation in our VT is accomplished using a two-level DW transformation which aligns signals both temporally (DTW) and spectrally (DFW).

Let $i = 1, 2$ denote the index of a speaker, with $i = 1$ representing the source speaker and $i = 2$ representing the target speaker. Let ℓ be the index of a segment of speech, where successive segments may be overlapping. Let $\mathbf{r}_i(\ell, :)$ denote the time-domain signal associated with the ℓth segment of speech. This is a vector of length $2K = 512$. This was transformed using a 512-point FFT. This was normalized as follows. Let $D_i(\ell, j)$, $i = 1, 2$, $\ell = 1, 2, \ldots, L_i$, $j = 1, 2, \ldots, K$ be the magnitude of the positive frequency coefficients in the DFT of $\mathbf{r}_i(\ell, :)$. The feature vector is produced by

$$\mathbf{s}_i(\ell, j) = \frac{D_i(\ell, j)}{\frac{1}{KL_i}\sqrt{\sum_{\ell=1}^{L_i}\sum_{j=1}^{K} D_i(\ell, j)^2}}, \qquad \ell = 1, 2, \ldots, L_i, \quad j = 1, 2, \ldots, K.$$

A sequence of these feature vectors is denoted as

$$\mathcal{S}_i = \{\mathbf{s}_i(1,:), \mathbf{s}_i(2,:), \ldots, \mathbf{s}_i(L_i,:)\}, \qquad i = 1, 2. \tag{1}$$

Here, L_1 and L_2 denote the number of source and target feature vectors. Let $d_T(\ell_1, \ell_2)$ denote the minimum cost between the sequences $\mathcal{S}_1$ (source) and $\mathcal{S}_2$ (target), up to segment indices ℓ_1 and ℓ_2, computed using DTW to align the segments in time. (The subscript T refers to time warping.) As usual for DTW [41], $d_T(\ell_1, \ell_2)$ is recursively computed as

$$d_T(\ell_1, \ell_2) = d(\mathbf{s}_1(\ell_1,:), \mathbf{s}_2(\ell_2,:)) + \min\{d_T(\ell_1 - 1, \ell_2), d_T(\ell_1, \ell_2 - 1), d_T(\ell_1 - 1, \ell_2 - 1)\}. \tag{2}$$

Here, $d(\mathbf{s}_1(\ell_1,:), \mathbf{s}_2(\ell_2,:))$ is the distance between individual feature vectors at segments ℓ_1 and ℓ_2. This is computed using DFW, as described below. Conceptually, the dynamic warping aligns the peaks and valleys between the source and target data, as suggested by Figure 2. When $\ell_1 = L_1$ and $\ell_2 = L_2$, the temporal warping process determines an overall warped metric distance $d_T(L_1, L_2)$ between the sequences $\mathcal{S}_1$ and $\mathcal{S}_2$. DTW computes a sequence of indices $\mathbf{a}_T = (a_T(1), a_T(2), \ldots, a_T(N_T))$ and $\mathbf{b}_T = (b_T(1), b_T(2), \ldots, b_T(N_T))$, which are called the temporal warping function paths. The temporal warping function paths define a time-aligned (TA) function $\mathbf{s}_{1,TA}$ such that the sequence

$$\mathbf{s}_{1,TA}(a_T(j),:) = \mathbf{s}_1(b_T(j),:), \quad j = 1, 2, \ldots, N_T$$

is as similar to $\mathcal{S}_2$ as possible.

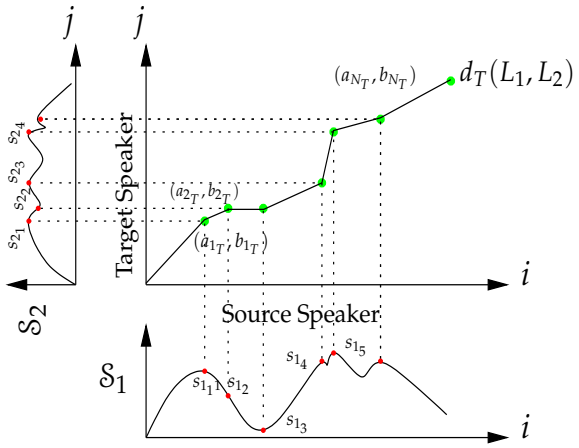


Figure 2. Distance measure for outer dynamic warping for VT.

The distance between two feature vectors in (2), $d(\mathbf{s}_1(\ell_1,:), \mathbf{s}_2(\ell_2,:))$, can itself be computed using warping between the components of these individual vectors. Because these vectors represent frequency information, this is DFW. DFW is computed in a way similar to the DTW. Let $\mathbf{s}_i = \mathbf{s}_i(\ell_i,:)$, $i = 1, 2$ denote the spectral information vectors at segment ℓ_i that is passed to the DFW function from the (outer) DTW function. The elements of $\mathbf{s}_i$ are denoted by $s_i(k)$. DFW is applied to calculate the distance between $\mathbf{s}_1$ and $\mathbf{s}_2$ as

$$d_F(k_1, k_2) = \text{dist}(s_1(k_1), s_2(k_2)) + \min\{d_F(k_1 - 1, k_2), d_F(k_1, k_2 - 1), d_F(k_1 - 1, k_2 - 1)\}, \tag{3}$$

where the subscript $_F$ refers to a frequency (or spectral) warping process, and $\text{dist}(s_1(k_1), s_2(k_2))$ represents the metric distance between elements of the feature spectral vectors.

The metric distance, $\text{dist}(s_1(k_1), s_2(k_2))$, used looks only at the magnitude of the spectral information, so

$$\text{dist}(s_1(k_1), s_2(k_2)) = (\,|\,|s_1(k_1)| - |s_2(k_2)|\,|\,). \tag{4}$$

When there are low energy speech segments (such as from unvoiced speech) which have no particular spectral information to match, the distance $d(\mathbf{s}_1, \mathbf{s}_2)$ is generally noninformative. In order to downplay this effect the small energy segments, the minimum energy $E_{\text{min}} = \min(\|\mathbf{s}_1\|^2, \|\mathbf{s}_2\|^2)$, is used to scale the distance, as follows. At the end of frequency warping process, the metric distance between spectral vectors to be used in (2) is computed as $d(\mathbf{s}_1, \mathbf{s}_2) = E_{min}\, d_F(K, K)$.

The DFW on each feature vector produces a sequence of indices $\mathbf{a}_F = (a_F(1), a_F(2), \ldots, a_F(N_F))$ and $\mathbf{b}_F = (b_F(1), b_F(2), \ldots, b_F(N_F))$. $\mathbf{a}_F$ and $\mathbf{b}_F$ are called the spectral warping function paths. Here N_F is the length of the spectral warping paths, which may be different from segment to segment.

The temporally aligned feature vector of speaker $\mathbf{s}_{1_{TA}}$ is transformed to spectrally match the target spectrum information of speaker $\mathbf{s}_2$ by creating a modified, spectrally-aligned (SA), feature vector according to

$$\hat{s}_{1,TF}(a_F(i)) = s_{1,TA}(b_F(i)), i = 1, 2, \ldots, N_F. \tag{5}$$

This SA map drags spectral components of source vectors to produce transformed voice spectra. Each time-aligned and spectrally-aligned segment is inverse Fourier transformed to produce the time signal for segment ℓ, $\hat{\mathbf{r}}_2(\ell, :)$. This time signal is windowed (we used a Hamming window of length 512 with 16-ms overlapping when the sample rate was 16,000 samples/sec) and the overlapping segments are added together.

The combination of the inner and outer warping processes applied to speech signal is portrayed in Figure 3. Starting at the bottom of the diagram, a speech signal is split into different overlapping segments. Spectral features are computed for each segment. These spectral feature vectors are passed through DTW for time alignment, and at every stage of DTW, the metric between vectors is computed using DFW. The method for doing two-level DW to achieve VT is given in Algorithm 1.

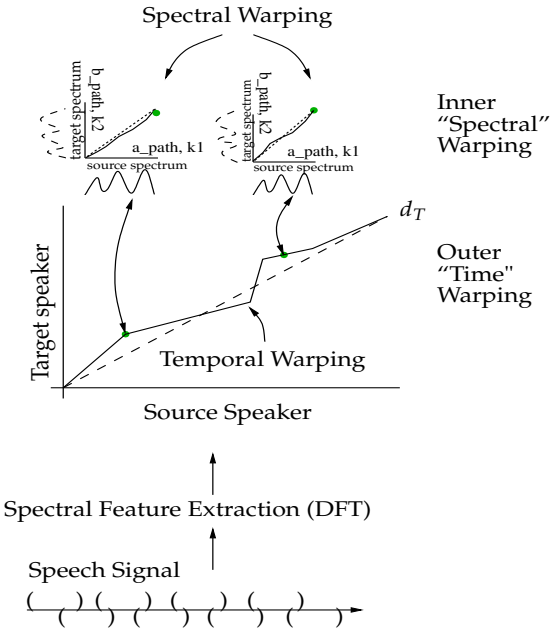


Figure 3. Inner/outer dynamic warping.

Algorithm 1: Dynamic Warping (DW) for Voice Transformation.

Input:
First spectral sequence, $\mathcal{S}_1 = \{\mathbf{s}_1(1, 1:n_T), \mathbf{s}_1(2, 1:n_T), \ldots, \mathbf{s}_1(n_F, 1:n_T)\}$
Second spectral sequence, $\mathcal{S}_2 = \{\mathbf{s}_2(1, 1:m_T), \mathbf{s}_1(2, 1:m_T), \ldots, \mathbf{s}_1(m_F, 1:m_T)\}$
window size w_T for DTW process
window size w_F for DFW process
Output:
Distance between $\mathcal{S}_1$ and $\mathcal{S}_2$
Indices $\mathbf{a}(N)$ and $\mathbf{b}(N)$ for temporal alignment
Indices $\mathbf{b}(M)$ and $\mathbf{b}(M)$ for spectral alignment
Begin
Initialize DTW array, $DTW = array[0 \ldots n_T, 0 \ldots m_T]$
Initialize DFW array, $DFW = array[0 \ldots n_S, 0 \ldots m_F]$
Adapt window size, $w_T = max(w_T, abs(n_T - m_T))$
For $i_T = 1$ *to* n_T
  For $j_T = 1$ *to* m_T
    $DTW[i_T, j_T] = \infty$
  End For j_T
End For i_T
Set $DTW[0, 0] = 0$
For $i_F = 1$ *to* n_F
  For $j_F = 1$ *to* m_F
    $DTW[i_F, j_F] = \infty$
  End For j_F
End For i_F
Set $DFW[0, 0] = 0$
For $i_T = 1$ *to* n_T
  For $j_T = max(1, i_T - w_T)$ *to* $min(m_T, i_T + w_T)$
    Adapt window size, $w_F = max(w_F, abs(n_S - m_F))$
    For $i_F = 1$ *to* n_F
      For $j_F = max(1, i_F - w_F)$ *to* $min(m_F, i_F + w_F)$
        $cost = abs(abs(\mathcal{S}_1(i_T, i_F))) - (abs(\mathcal{S}_2(j_T, j_F)))$
        $DFW[i_F, j_F] := cost + minimum(DTW[i_F - 1, j_F],$
                                $DFW[i_F, j_F - 1],$
                                $DFW[i_F - 1, j_F - 1])$
      End For j_F
    End For i_F
    Searching minimum path through $DFW[i_F, j_F]$, save $\mathbf{a}_F$, $\mathbf{b}_F$
    $cost_{new} = DFW[n_F, m_F]$
    Minimum energy [$\mathbf{s}_1$ and $\mathbf{s}_2$]
    $d = E_{min}\ cost_{new}$
    $DTW[i_T, j_T] := d + minimum(DTW[i_T - 1, j_T],$
                            $DTW[i_T, j_T - 1],$
                            $DTW[i_T - 1, j_T - 1])$
End For j_T
End For i_T
Searching minimum path through $DTW[i_T, j_T]$, save $\mathbf{a}_T$, $\mathbf{b}_T$

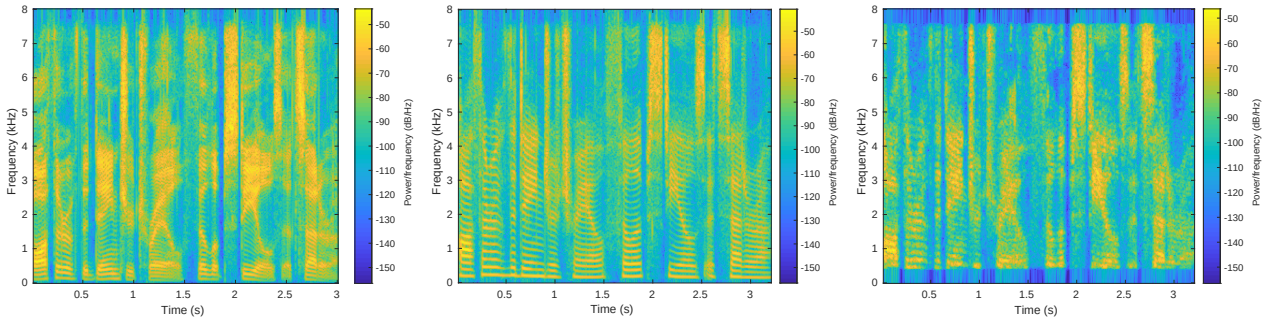
3.2. Experiment Description and a Proof of Concept Result

Experiments were carried out on CMU ARCTIC database [42]. The database consisted of seven primary sets of recordings, recorded by two US males, one Canadian male with English accent, one Scottish male with English accent, one Indian male, and two US females. Each speaker recorded a set of 1132 English utterances, most being between one and four seconds long and between 5 to 15 words long. The prompt list was designed to be

phonetically balanced American English and have diphone coverage representative of the source material. The sound files were recorded in a soundproof booth at 16,000 samples/sec with simultaneous EGG measurements (the speech needed to be clean, or else background disturbances affected the transformed speech). The lexical and phonetic descriptions derived from the US English front-end module. In experiments presented here, one US male and one US female were chosen to do VT from male to female. While this was a limited experiment, since the transformation process was general (being based on DW of spectral features), this was considered a valid proof of concept for this initial experiment.

After identifying the starting point and ending point of the first 720 sentences of CMU ARCTIC database, each sentence was temporally segmented into 32-ms segments, using a Hamming window with 16-ms overlapping, zero-padded, then transformed using a 512-point FFT then normalized as described above. DW was applied on these features and the features of the first 600 sentences were used to train the NN and a separate set of 120 sentences were used for testing the trained NN.

Figure 4a shows a typical spectrogram for a male speaker, using the phrase "Author of the danger trail, Philip Steels, etc." from the CMU ARCTIC database. Figure 4b shows the spectrogram for a female speaker for the same phrase. Figure 4c shows the VT from male speaker to female speaker. After transformation, the overall energy distribution looked similar to the spectrogram of the female speaker. The fine-scale variations (the ripples in frequency) evident in Figure 4a,b are due to pitch. These lines are also in Figure 4c, but the pitch details are fuzzier. The resulting speech had very severe processing artifacts. (It did not sound very good.) Since the transformation only involved shifting magnitude spectral information, we were led to suspect that the problem might result from a mismatch between the magnitude and phase information of the warped signal.



(**a**) Male Spectrogram Information (**b**) Female Spectrogram Information (**c**) Spectrogram of warped Male to Female

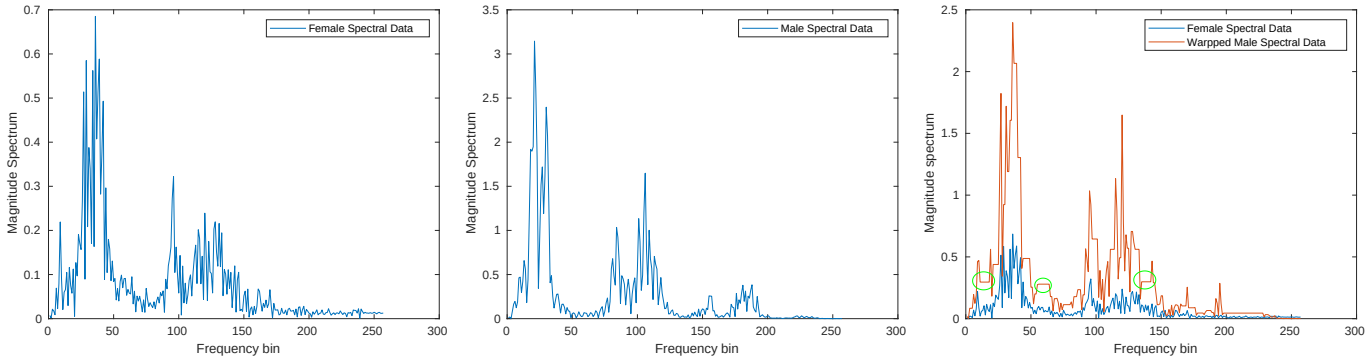
Figure 4. Spectrogram for male, female and warped male to female.

Figure 5a shows the spectral feature information for one segment of speech for the female speaker. Figure 5b shows the spectral feature information for one time-aligned segment of speech for the male speaker. By applying the DFW, the peaks and valleys of the male spectrum were generally aligned to the peaks and valleys of the female spectrum, as portrayed in Figure 5c. In warping the spectral information, artifacts were introduced into the signal which were not typical of speech. For example, Figure 5c shows that the warping smeared some spectral values across several frequency bins (indicated with green circles) in the plot, resulting in regions of the spectrum that had constant spectral magnitude.

We found that the quality of the speech was significantly improved by using a phase reconstruction algorithm on the warped speech. In this work, the Griffin–Lim algorithm [43] was applied to do phase reconstruction after the spectral transformation has been accomplished [44].

Figure 6 shows the spectrogram for the warped speech after phase reconstruction. The pitch lines were stronger (less fuzzy) after the GLA phase reconstruction. Figure 7 shows the spectral feature information for one segment of speech for the female speaker with the time aligned spectral warped segment for male from the phrase mentioned above.

The peaks and valleys were still aligned, but the regions where there were spectral flats in Figure 5c, denoted by the green circles, were replaced with more speech-like spectra.



(**a**) One Segment of Female Spectral Feature (**b**) One Segment of Male Spectral Feature (**c**) Warped Spectral Male with Female Spectral Feature

Figure 5. One spectral segment feature for male, female and warped male to female.

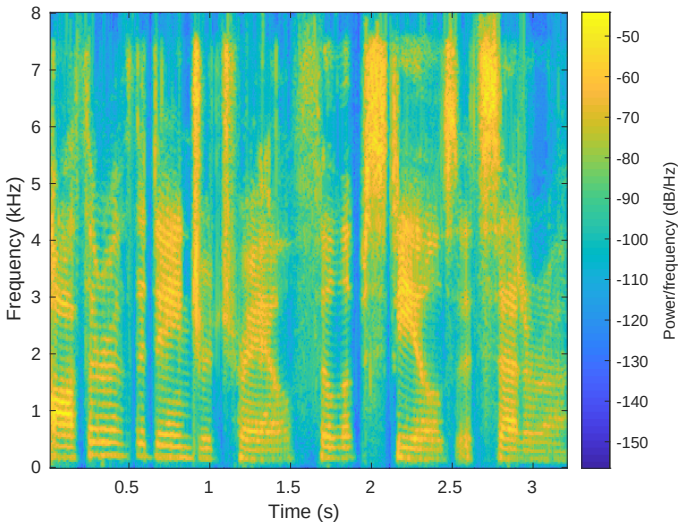


Figure 6. Spectrogram information for warped male to female with using GLA.

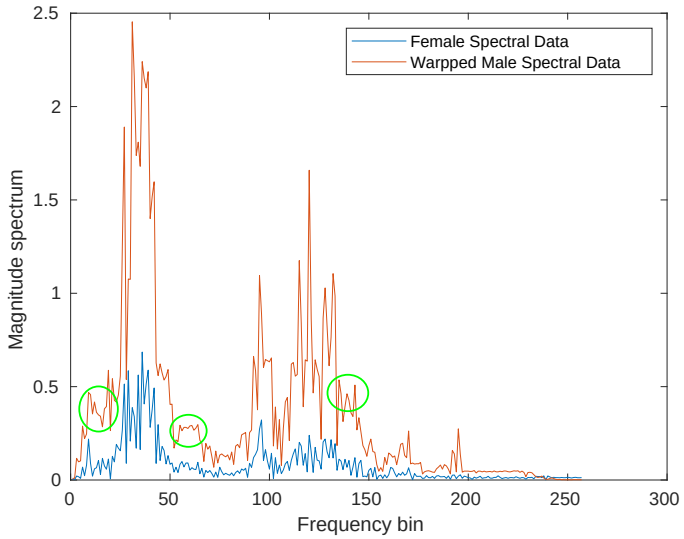


Figure 7. One spectral segment feature for warped male to female with GLA.

Table 1 shows the MCD score obtained using warping with and without the GLA phase reconstruction. The MCD was nearly 3 dB better when GLA phase reconstruction was used. The audio file also sounded significantly better. This verified that the MCD was valid as a measure of audio quality, and that the phase reconstruction was an important part of the VT. Phase reconstruction was used in all transformation experiments described below.

Table 1. MCDs obtained for 2-Level DW with and without GLA.

No.	Two-Level DW Method	MCD [dB]
1	DW, with phase reconstruction	2.4
2	DW, without phase reconstruction	6.3

4. Spectral Warping Using NN

The proof-of-concept experiments above demonstrated that spectral mapping followed by phase reconstruction produces effective VT. For a complete, viable VT system, it is necessary to establish conversion rules for transforming speech when the target speaker saying the desired phrase is not available. These conversion rules are represented using a NN which is trained to accept spectral inputs from the source speech and to produce the spectral warping function paths, $\mathbf{a}_F$ and $\mathbf{b}_F$. The process is outlined in Figure 8. The input data for training the NN is the magnitude of the spectrum of a segment of speech that has been time-aligned with the target speech. The spectral information for segments of speech from source and target are used in DW to find the $(\mathbf{a}_F, \mathbf{b}_F)$ spectral warping information, as suggested by Figure 8a. The NN is trained to produce this $(\mathbf{a}_F(i), \mathbf{b}_F(i))$ information when the magnitude spectrum for the segment $\mathbf{s}_{1_{TA}}$ is the input. This is suggested by the NN diagram in Figure 8b.

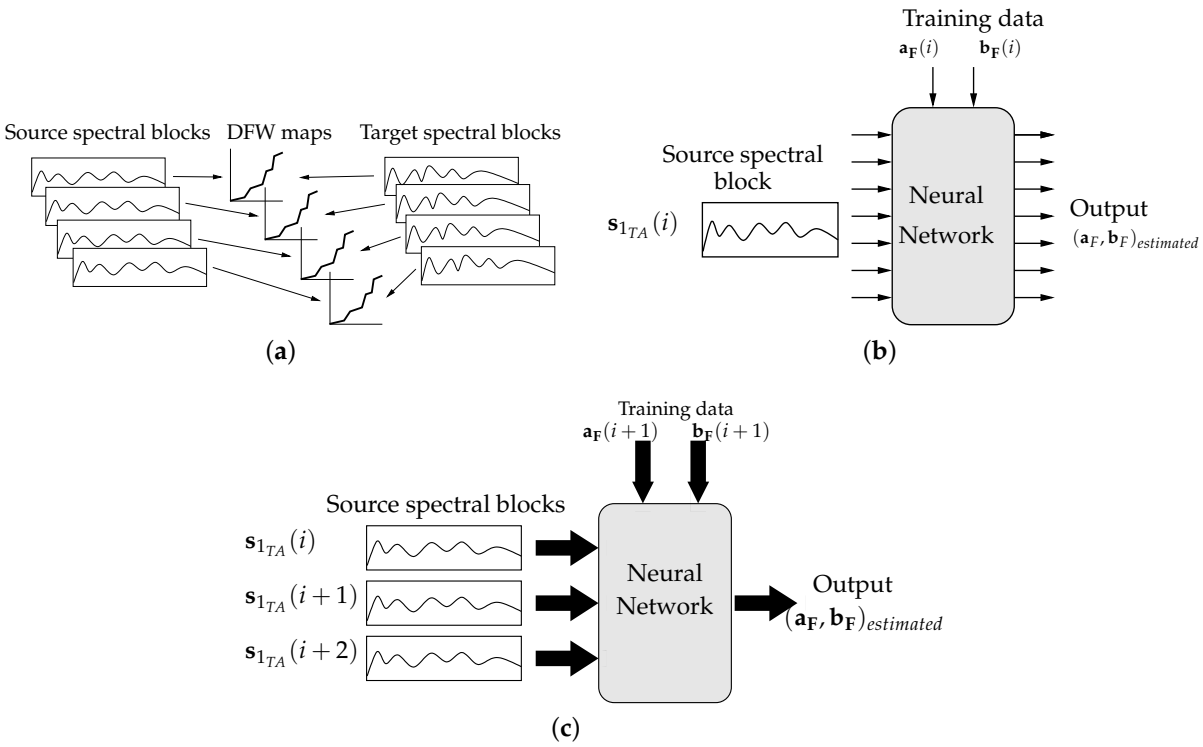


Figure 8. Spectrogram for male, female and warped male to female. (**a**) Making DFW training data. (**b**) Training a neural network—one input spectral vector. (**c**) Training a neural network—three input spectral vectors.

It was found that better VT was produced when the NN uses the spectral magnitude from the preceding segment $\mathbf{s}_{1_{TA}}(i)$ and the following segment $\mathbf{s}_{1_{TA}}(i+1)$ as inputs,

to provide context for the current segment $\mathbf{s}_{1_{TA}}(i+2)$. This configuration is suggested in Figure 8c.

Interpolating the Data to Achieve Constant Length

As noted above, the length of the warping path $(\mathbf{a}_F, \mathbf{b}_F)$ may vary among the different segments of speech. This varying length poses a problem for a NN which has a fixed number of outputs. Rather than zero-pad the outputs of the NN to produce a desired length (which would create artifacts) the $\mathbf{a}_F$ and $\mathbf{b}_F$ are interpolated to produce a modified $\mathbf{a}_F$ and $\mathbf{b}_F$ which is the same (maximum) length for all segments of the speech vector. In our experiments the maximum length of the warping paths was 482 (as determined from training data). The interpolation is described using Matlab-like notation and functions. Let $\max_{\text{index}}$ denote the maximum length of the interpolated vectors. The interpolation procedes in two steps. First, interpolated indices are computed according to

$$\tilde{n} = [1 : \max_{\text{index}}] \times \frac{\text{length}(\mathbf{a})}{\max_{\texttt{index}}};$$

Then interpolation among the $\mathbf{a}$ values using these interpolated indices is computed by

$$\tilde{\mathbf{a}} = \text{interpolate}(\mathbf{a}, [1 : \text{length}(\mathbf{a})], \tilde{n})$$

The sequence $\mathbf{b}$ is similarly interpolated. Figure 9 illustrates un-interpolated function paths for $\mathbf{a}$ and $\mathbf{b}$ with interpolated paths ($\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$ respectively). The interpolated $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$ have the same general shape as the original $\mathbf{a}$ and $\mathbf{b}$, but have length 482, as desired. The NN is trained to produce the interpolated warping sequences at its output.

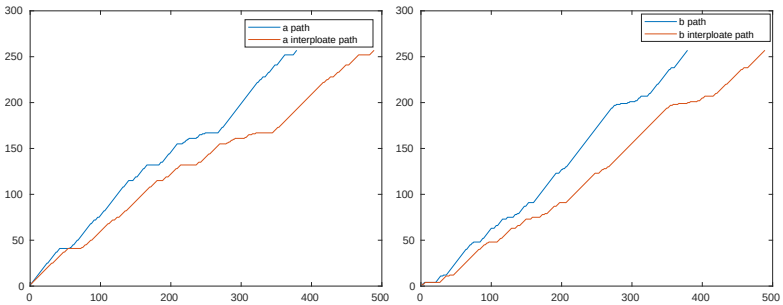


Figure 9. Un-interpolated and interpolated paths. (**a**) $\mathbf{a}$ and $\tilde{\mathbf{a}}$. (**b**) $\mathbf{b}$ and $\tilde{\mathbf{b}}$.

5. Mel-Cepstral Distortion as an Objective Measure

We use Mel-Cepstral Distortion (MCD) to evaluate the quality of the transformed speech, and to be able to compare our work to other work in the literature. MCD has been used as an objective error measure for evaluating the quality of synthetic voice [10]. It is a measure of the difference between two sequences of mel-cepstra. It is computed as

$$MCD = (10/\ln 10)\sqrt{2\sum_{i=1}^{25}(mc_i^{(t)} - mc_i^{(w)})^2} \tag{6}$$

where $mc_i^{(t)}$ and $mc_i^{(w)}$ represent the mel-cepstrum of the target speech and warped speech, respectively.

We used MCD both as a way of evaluating the performance of different NN architectures, as well as a tool for comparing with previous VT research which reports MCD scores.

6. NN Architecture Experiments

A series of experiments was conducted to determine an effective NN structure. Two different input models were considered. In the first model, a single input block of time-aligned spectral magnitude data, $\mathbf{s}_{1,TA}(:,i)$, is presented as the input of the NN. There are 256 inputs in the block (corresponding to the positive frequencies of the DFT of a 512-sample segment of speech). The NN has 482 outputs for each element of **a** and **b** for a total of 964 outputs. (482 is the largest length of the **a**, **b** data that was found among the training data.) This architecture is referred to as the 1-IP/1-OP model.

In the second model, three consecutive time-aligned vectors $(\mathbf{s}_{1,TA}(:,i), \mathbf{s}_{1,TA}(:,i+1), \mathbf{s}_{1,TA}(:,i+2))$ were used as the input, for a total of 768 inputs. The three inputs provided temporal/frequency context for the central block. There were still 964 outputs. This architecture is called the 3-IP/1-OP model.

Training was performed by Tensorflow and Keras to estimate the warping paths. The neural architecture was developed in a series of phases, as described below. Initialize weights were selected at random. Each optimizer was trained for different learning rates (1, 0.1, 0.01, 0.001, 0.0001 and 0.00001); we found that a learning rate of 0.001 achieved the best results. Our NN models were trained with different optimizers (Gradient Descent, Adam and RMS Prop). The best optimizer for this work was the Gradient Descent. The NN architecture was developed in a series of phases, as described below.

6.1. Phase One

This phase was designed as a proof of concept: can a NN learn the spectral warping function paths? For this phase, a single phrase was selected to be trained using the two input models mentioned above. This phase also provided a basis for exploring different NN architectures. For all these tests, the phrase, "Author of the danger trail, Philip Steels, etc." was selected from the CMU-ARCTIC database.

We experimented with both input models on different number of layers in order determine a good architecture. Table 2 shows the number of nodes in each layer and the output function used for that layer for both architecture. The notation 256L 500R 900R 1000R 964L means that this NN had 5-layers with 256 inputs, 964 output nodes with 500, 900, and 1000 neuron nodes in the hidden layers. L and R represent "linear" and "ReLU" activation output functions, respectively. This table also shows results for architectures described in later phases of development.

Figure 10 shows how the NN learned warping paths at various segments of the selected speech signal for different training iterations. In this figure, each subplot represents the mapping for one spectral segment of speech. In each of these, the blue plots shows the "truth" from DW the paths; the NN-learned paths are in orange. In this figure, the horizontal axes are the sequence of indices (1–468) and the vertical axes are the values of the $\mathbf{a}_F$, truth and learned. Figure 10a uses the 19-layers-1-IP/1-OP architecture. There was fairly good matching between the training data and the NN output, but Figure 10b, produced using the 19-layers-3-IP/1-OP architecture showed better results, with the training data being virtually indistinguishable from the NN outputs. Figure 10c,d shows how the NN learned a **b** warping path at various segments of the selected speech signal after 5000 training iterations. Again an excellent match was achieved with 3-IP/1-Op architecture. Not only was the NN better trained as demonstrated by these figures, but the quality of the audible transformed signal was significantly better when the three-input NN was used. Figure 11 shows the MSE between the true and estimated warping paths for the selected architecture (3-IP/1-OP) with 5000 training iterations, demonstrating that the learning essentially converged by that point.

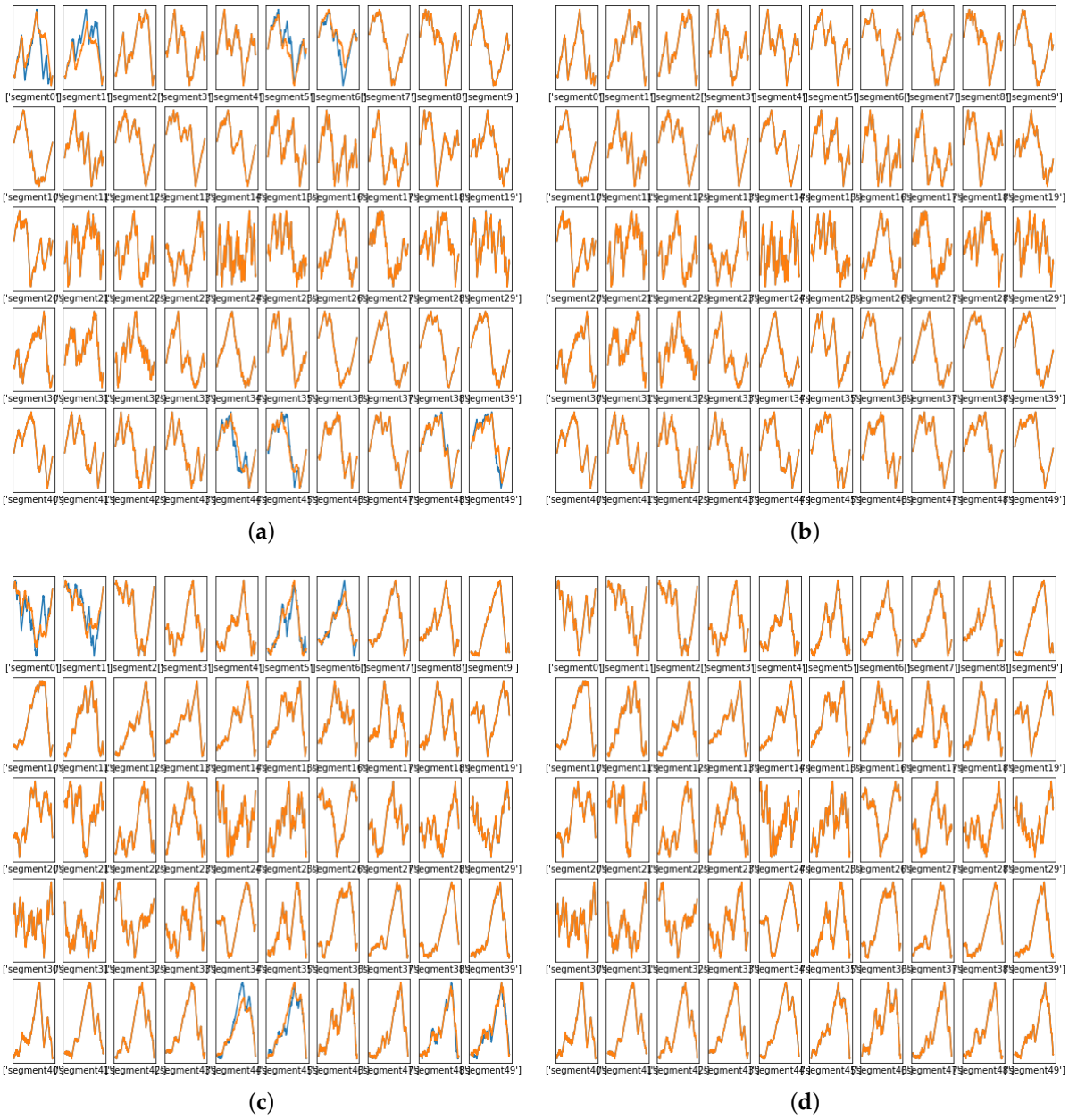


Figure 10. Learned warping paths, phase one, 5000 training iterations. (**a**) Warping path **a** produced using 1-IP/1-OP architecture. (**b**) Warping path **a** produced using 3-IP/1-OP architecture. (**c**) Warping path **b** produced using 1-IP/-1-OP architecture. (**d**) Warping path **b** produced using 3-IP/1-OP architecture.

The use of three input blocks provided results superior to the single input block model. On the basis of the MCD values from Table 2 and Figure 10b,d, we concluded that the NN used in the phases of this experiment should be based on the 19-layers-3-IP/1-OP architecture.

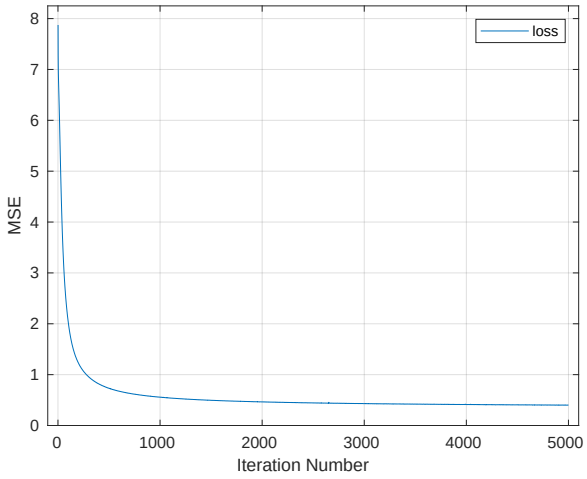


Figure 11. Mean squared error (l curve), phase one.

Table 2. MCDs obtained for different architectures.

Phase No.	Input Model	Architecture Type	NN Architecture	MCD [dB]
Phase One	1-IP-1-OP	5	256L 600R 1200R 1000R 964L	20.5
		6	256L 600R 1000R 1500R 1000R 964L	18.01
		7	256L 600R 1000R 1500R 2000R 1500R 964L	18.002
		19	500L 1500R 2500R 3500R 4500R 5040R 6459R 7459R 8459R 9459R 10459R 8459R 7939R 6939R 5500R 4500R 3500R 2500R 964L	3.3
		20	500L 1500R 2500R 3500R 4500R 5040R 6459R 7459R 8459R 9459R 10459R 9459R 8459R 7939R 6939R 5500R 4500R 3500R 2500R 964L	5.5
	3-IP-1-OP	5	768L 4500R 6500R 6939R 964L	16.85
		6	768L 4500R 6500R 7040R 6939R 964L	15.3
		7	768L 4500R 6500R 8040R 8939R 6939R 964L	12.92
		9	768L 4500R 6500R 7040R 9459R 8939R 7939R 6939R 964L	7.5
		10	768L 4500R 6500R 7040R 9459R 10459R 8939R 7939R 6939R 964L	4.7
		19	**1500L 2500R 3500R 4500R 6500R 7040R 8459R 9459R 10459R 10459R 9459R 8459R 7939R 6939R 5500R 4500R** 3500R 2500R 964L	**2.25**
		20	1500L 2500R 3500R 4500R 6500R 7040R 8459R 9459R 10459R 11459R 10459R 9459R 8459R 7939R 6939R 5500R 4500R 3500R 2500R 964L	3.5
Phase Two	3-IP-1-OP	19	1500L 2500R 3500R 4500R 6500R 7040R 8459R 9459R 10459R 9459R 8459R 7939R 6939R 5500R 4500R 3500R 2500R 964L	4.8
Phase Three	3-IP-1-OP	Six-Clusters.Each Cluster trained for 19 layers	1500L 2500R 3500R 4500R 6500R 7040R 8459R 9459R 10459R 10459R 9459R 8459R 7939R 6939R 5500R 4500R 3500R 2500R 964L	2.8
Phase Four	1-IP-1-OP	7	256L 64R (k = 3) 100R (k = 3) 150R (k = 3) 100R (k = 3) FL D100R 964L	13.33
	3-IP-1-OP	5	768L 64R (k = 3) 100R (k = 3) FL D100R 964L	13.5

6.2. Phase Two

After validating from Phase One that the NN was able to produce good transformed speech for a single phrase, the NN was trained to deal with multiple phrases. In this phase, training was conducted using 600 phrases from a single speaker for training, and a separate set of 120 phrases from a single speaker was used to test. The selected NN architecture was trained for 10,000 iterations.

Figure 12 shows the warping paths for some of the segments, where the DW path (truth) is shown in blue and the learned path is in orange. The NN learned some of the

warping paths well. Figure 12b, which was produced using the selected architecture after 10,000 training iterations, showed more matching between the $\mathbf{a}_F$ warping path and the learned path than Figure 12a, which was produced at 2500 training iterations. Figure 12d, which was produced at 10,000 training iterations, showed more matching between the $\mathbf{b}_F$ warping path and the learned path than Figure 12c, which produced at 2500 training iterations. By 10,000 training iterations, many of the warping functions were learned well, but there was still a significant number that were not learned well. Figure 13 shows the MSE between the true values of the warping paths and the estimated one.

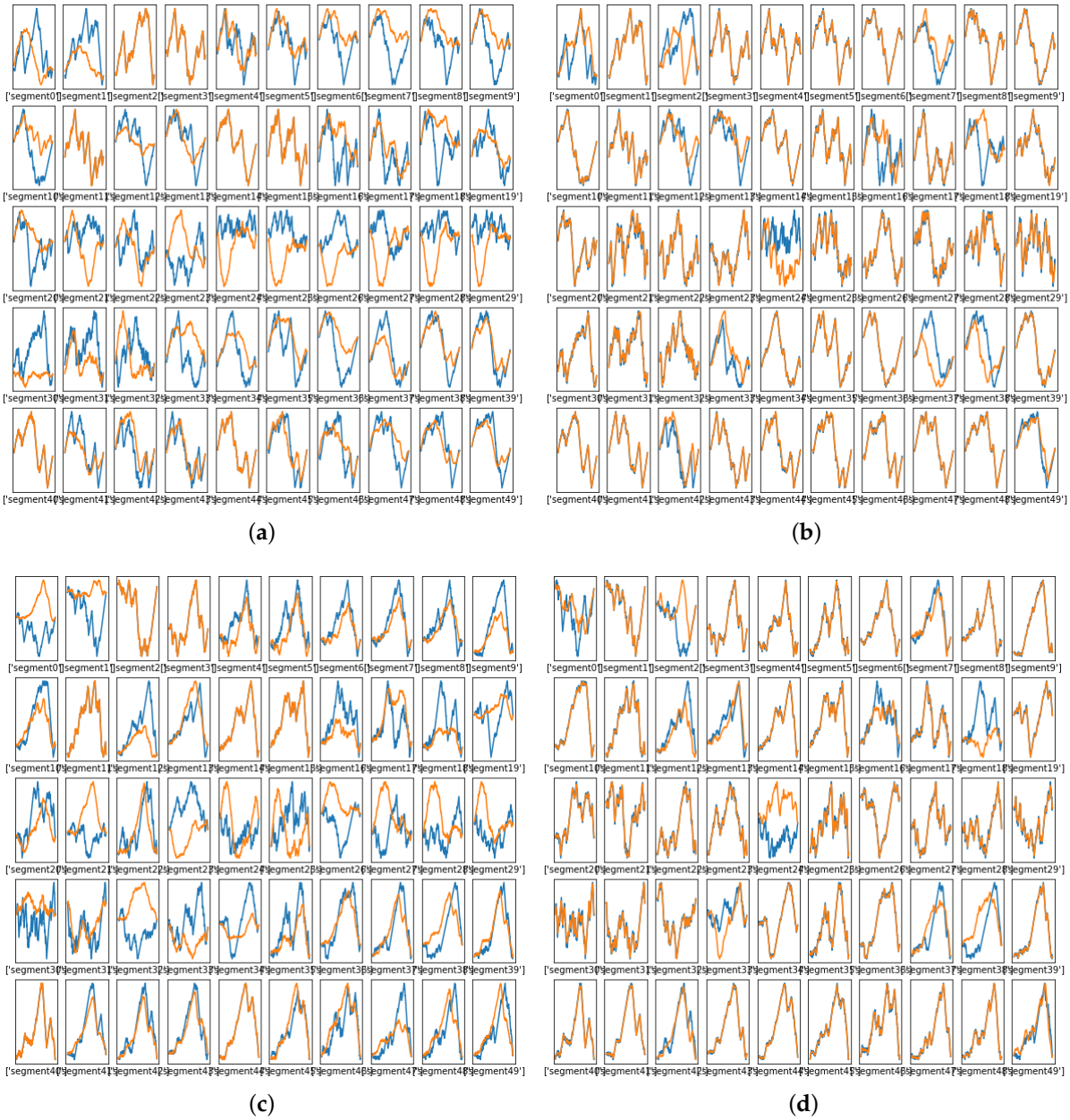


Figure 12. Learned warping path **a**, Phase Two. (**a**) Warping path **a**, 2500 iterations. (**b**) Warping path **a**, 10,000 iterations. (**c**) Warping path **b**, 2500 iterations (**d**) Warping path **b**, 10,000 iterations.

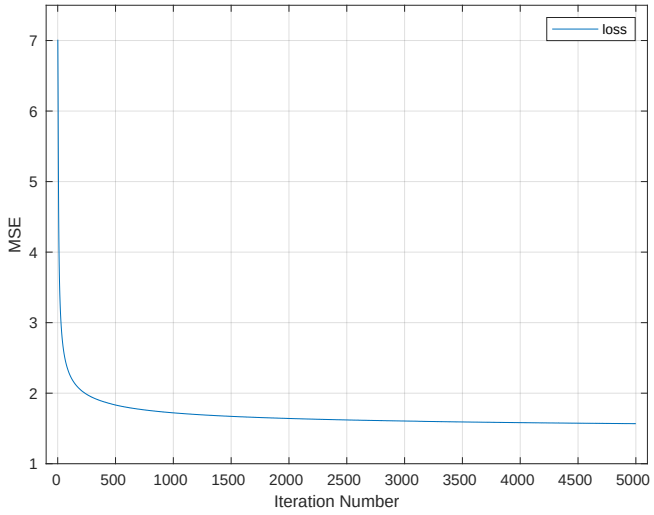


Figure 13. Mean squared error learning curve, Phase Two.

6.3. Phase Three

While the results of Phase Two were promising, there was still significant error in the estimates of the warping functions. To further improve the performance, it was determined to assist the NN by decomposing the problem into smaller pieces by clustering. To this end, several NNs were trained, each to be responsive to data pertaining to a particular cluster, where clustering was accomplished using k-means clustering. Different numbers of clusters were tried: $k = 4, 5, 6, 10, 15, 20$. It was found that $k = 6$ clusters gave good performance. Figure 14 shows the $k = 6$ clusters for the selected spectral data from the 600 training phrases. The six clusters are shown using different colors.

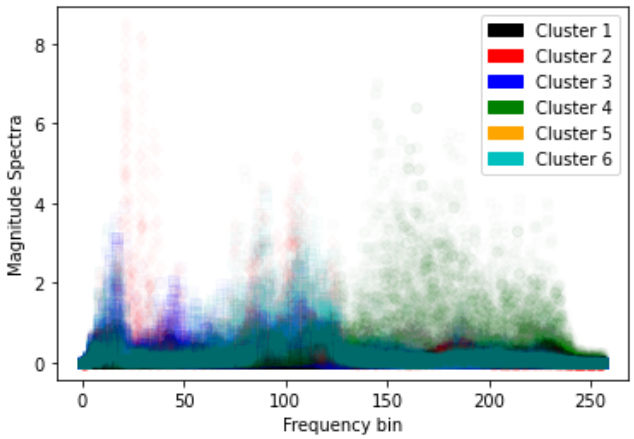
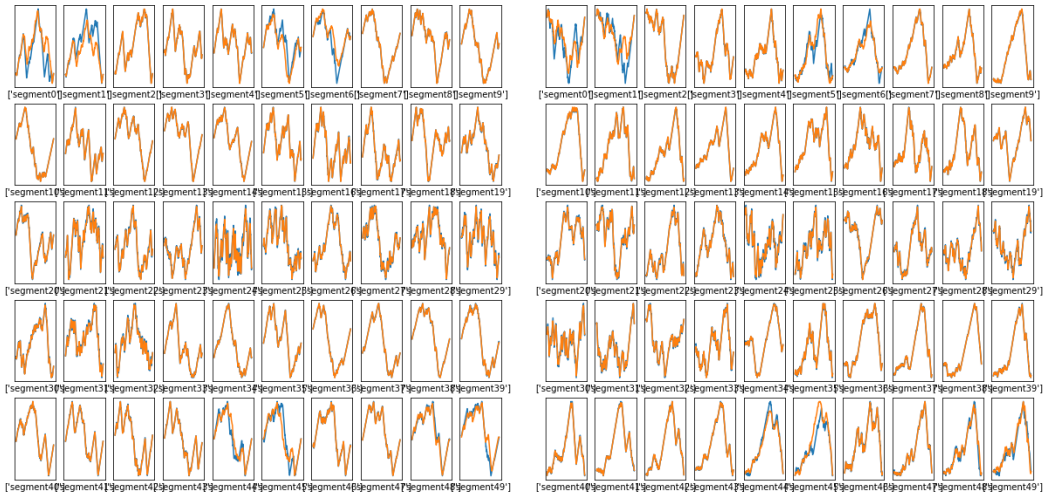


Figure 14. Clustering of spectral information from 600 phrases ($k = 6$).

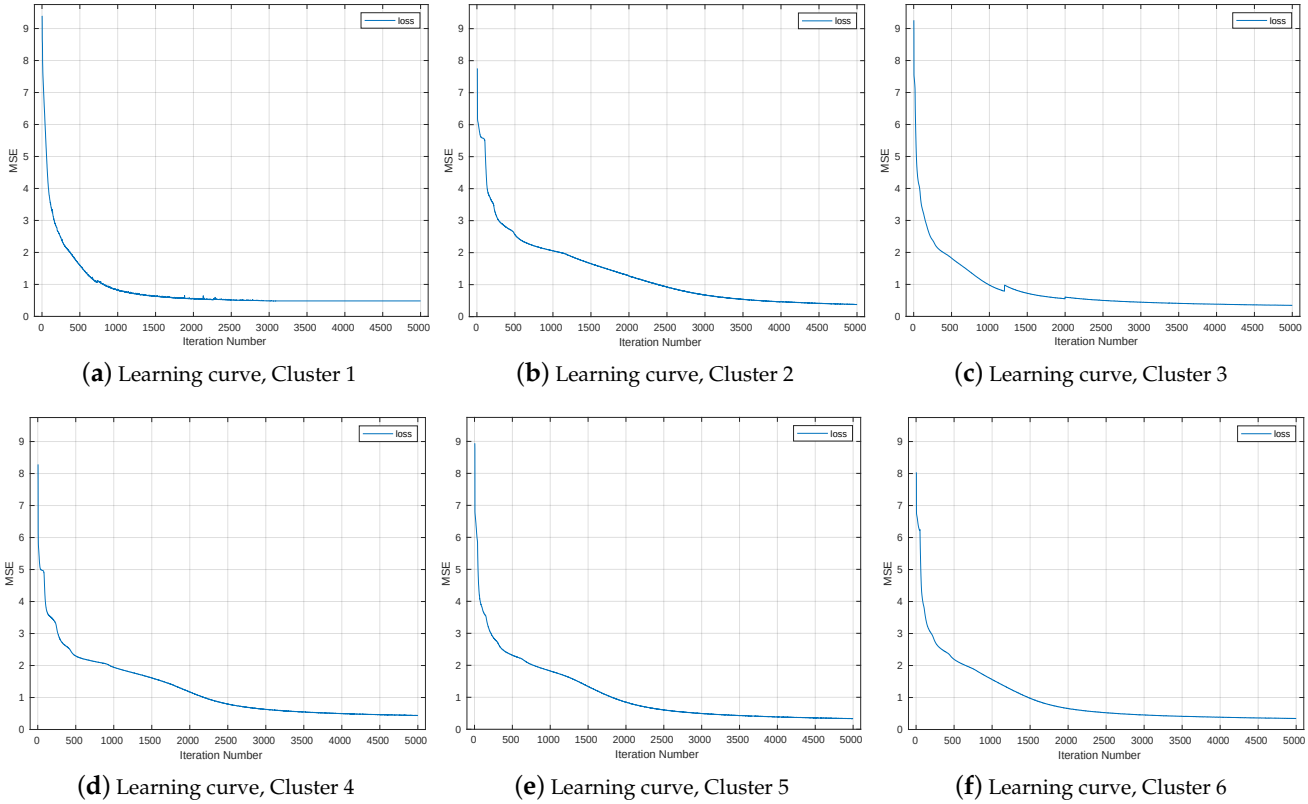
In the training stage, data from each of the k clusters were used to train k NNs. In the synthesis stage, for each segment of speech data it was determined which cluster the associated feature vector best matched, and the $(\mathbf{a}_F, \mathbf{b}_F)$ warping functions were obtained from the corresponding NN. These were used to provide the spectral warping.

Figure 15a,b show how the clustering NN learned the paths **a** and **b**, respectively, at 5000 training iterations. The DW path (truth) is shown in blue and the NN-learned path is in orange. At this point, the final sound at the 5000 training iterations is much better than the final sound produced from Phase Two, with little signal processing artifacts. Figure 16 shows the MSE between the true values of the warping paths (**a** and **b**) and the estimated paths for each of the six clusters. Interestingly, the learning demonstrated by the MSE varied from cluster to cluster. The MCD value for this phase is reported in Phase Three row of Table 2.



(**a**) (**b**)

Figure 15. Learned warping paths **a** and **b**, Phase Three. (**a**) Warping path **a**, 5000 iterations. (**b**) Warping path **b**, 5000 iterations.



(**a**) Learning curve, Cluster 1 (**b**) Learning curve, Cluster 2 (**c**) Learning curve, Cluster 3

(**d**) Learning curve, Cluster 4 (**e**) Learning curve, Cluster 5 (**f**) Learning curve, Cluster 6

Figure 16. Mean squared error (learning curve) using cluster method.

6.4. Other NN Structures

The fully connected NN structures described above were shown to produce the warping functions with good fidelity. However, they took a lot of training and a lot of memory. In the interest of reducing the NN complexity, we examined convolutional NN structures (CNNs). However, CNNs failed to produce good acoustic results. The MCD scores for CNN architectures are reported in Table 2 as Phase Four. The CNN structures performed significantly worse than the fully connected NN structures.

7. Comparison

From Table 2 the average MCD across the test data was MCD = 2.8 dB. By contrast, in [32], the MCD for male to female conversion was 5.5 dB. The performance of our method was better than reported in [5], which used NN and reported an average MCD of 6.55. The work of VT using NN in [31] reported MCD of 6.1 dB. In all comparisons made, our method was significantly better (as measured by MCD) than previous work. Table 3 shows more comparison between this work and other works, mentioned in the references part, that used MCD scores to evaluate their works. The method presented here worked better than all previous work that reported performance using MCD.

Table 3. Comparative MCDs from previous work.

Reference Number	Reference Year	MCD [dB]
This Works	2020	2.8
[30]	2002	5.6
[32]	2015	5.9
[33]	2016	5.1
[16]	2019	5.09
[29]	2020	4.11

8. Summary and Conclusions

VT has been achieved by employing two-level DW in combination with NN structures. The DW does both temporal and spectral alignment. The path warping functions from DFW are used to train a NN to produce the warping functions. These warping functions were interpolated to provide vectors of constant length. The warped amplitude spectra are shaped by phase reconstruction, which was shown to dramatically improve the speech quality.

To provide sufficient representation capability on larger training data sets, we found it necessary to cluster the training data, training each segment separately then combining the results. This produced low error in the representation of the warping functions.

Based on comparisons with mel-cepstral scores, the VT approach provides performance competitive with previous works.

The primary focus of this work has been on the development of the two-level DW and design of a NN architecture to do VT. Future work will consider transformation among different classes of speakers, including dealing with speakers with different accents.

Author Contributions: Conceptualization, A.-W.A.-D., T.K.M., J.H.G.; methodology, A.-W.A.-D., T.K.M.; software, A.-W.A.-D.; validation, A.-W.A.-D.; formal analysis, A.-W.A.-D.; investigation, A.-W.A.-D.; resources, T.K.M., J.H.G.; writing—original draft preparation, A.-W.A.-D.; writing—review and editing, T.K.M., J.H.G.; supervision, T.K.M.; project administration, T.K.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code, data, and transformed sounds can be found here: https://digitalcommons.usu.edu/etd/8030, https://github.com/AlwaledAldulaimi/Voice-Transformation-.git (accessed on 23 May 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Helander, E.; Schwarz, J.; Nurminen, J.; Silen, H.; Gabbouj, M. On the impact of alignment on voice conversion performance. In Proceedings of the 2008 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, 22–26 September 2008; Volume 1, pp. 1453–1456.

2. Narendranath, M.; Murthy, H.A.; Rajendran, S.; Yegnanarayana, B. Transformation of formants for voice conversion using artificial neural networks. *Speech Commun.* **1995**, *16*, 207–216. [CrossRef]
3. Turk, O.; Buyuk, O.; Haznedaroglu, A.; Arslan, L.M. Application of voice conversion for cross-language rap singing transformation. In Proceedings of the IEEE 2009 International Conference on Acoustics, Speech and Signal Processing, (ICASSP), Taipei, Taiwan, 19–24 April 2009; pp. 3597–3600.
4. Stylianou, Y. Voice transformation: a survey. In Proceedings of the IEEE 2009 International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taipei, Taiwan, 19–24 April 2009; pp. 3585–3588.
5. Ramos, M.V. Voice Conversion with Deep Learning. Masters of Science Thesis, Universidade do Lisboa, Lisboa, Portugal, 2016.
6. Abe, M.; Nakamura, S.; Shikano, K.; Kuwabara, H. Voice conversion through vector quantization. *J. Acoust. Soc. Jpn.* **1990**, *11*, 71–76. [CrossRef]
7. Kuwabara, H. Quality control of speech by modifying formant frequencies and bandwidth. In Proceedings of the 11th International Congress of Phonetic Sciences (ICPhS), Tallinn, Estonia, 1–7 August 1987; pp. 281–284.
8. Childers, D.; Yegnanarayana, B.; Wu, K. Voice conversion: Factors responsible for quality. In Proceedings of the 1985 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Tampa, FL, USA, 26 March 1985; Volume 10, pp. 748–751.
9. Valbret, H.; Moulines, E.; Tubach, J.P. Voice transformation using PSOLA technique. *Speech Commun.* **1992**, *11*, 175–187. [CrossRef]
10. Toth, A.R.; Black, A.W. Using articulatory position data in voice transformation. In Proceedings of the 6th International Speech and Communication Association (ISCA) Workshop on Speech Synthesis, Antwerp, Belgium, 27–31 August 2007; pp. 182–187.
11. Helander, E.; Virtanen, T.; Nurminen, J.; Gabbouj, M. Voice conversion using partial least squares regression. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 912–921. [CrossRef]
12. Benisty, H.; Malah, D. Voice conversion using GMM with enhanced global variance. In Proceedings of the 2011 12th Annual Conference of the International Speech Communication Association (ISCA), Florence, Italy, 28–31 August 2011.
13. Sathiarekha, K.; Kumaresan, S. A survey on the evolution of various voice conversion techniques. In Proceedings of the 2016 IEEE 3rd International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 22–23 January 2016; Volume 1, pp. 1–5.
14. Yang, Y.; Uchida, H.; Saito, D.; Minematsu, N. Voice Conversion Based on Matrix Variate Gaussian Mixture Model Using Multiple Frame Features. In Proceedings of the 2016 17th Annual Conference of the International Speech Communication Association (ISCA), San Francisco, CA, USA, 8–12 September 2016.
15. Tian, X.; Lee, S.W.; Wu, Z.; Chng, E.S.; Li, H. An exemplar-based approach to frequency warping for voice conversion. *IEEE Trans. Audio Speech Lang. Process.* **2017**, *25*, 1863–1876. [CrossRef]
16. Sisman, B.; Zhang, M.; Li, H. Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion. *IEEE Trans. Audio Speech Lang. Process.* **2019**, *27*, 1085–1097. [CrossRef]
17. Sisman, B.; Yamagishi, J.; King, S.; Li, H. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE Trans. Audio Speech Lang. Process.* **2020** *29*, 132–157. [CrossRef]
18. Toda, T.; Black, A.W.; Tokuda, K. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 2222–2235. [CrossRef]
19. Zen, H.; Nankaku, Y.; Tokuda, K. Probabilistic feature mapping based on trajectory HMMs. In Proceedings of the 2008 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, 22–26 September 2008.
20. Pilkington, N.C.; Zen, H.; Gales, M.J. Gaussian process experts for voice conversion. In Proceedings of the 2011 12th Annual Conference of the International Speech Communication Association, (ISCA), Florence, Italy, 28–31 August 2011.
21. Xu, N.; Tang, Y.; Bao, J.; Jiang, A.; Liu, X.; Yang, Z. Voice conversion based on Gaussian processes by coherent and asymmetric training with limited training data. *Speech Commun.* **2014**, *58*, 124–138. [CrossRef]
22. Takamichi, S.; Toda, T.; Black, A.W.; Nakamura, S. Modulation spectrum-based post-filter for GMM-based voice conversion. In Proceedings of the IEEE Signal and Information Processing Association Annual Summit and Conference (APSIPA), Chiang Mai, Thailand, 9–12 December 2014; pp. 1–4.
23. Takamichi, S.; Toda, T.; Black, A.W.; Neubig, G.; Sakti, S.; Nakamura, S. Postfilters to modify the modulation spectrum for statistical parametric speech synthesis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 755–767. [CrossRef]
24. Toda, T.; Muramatsu, T.; Banno, H. Implementation of computationally efficient real-time voice conversion. In Proceedings of the 2012 13th Annual Conference of the International Speech Communication Association, Portland, OR, USA, 9–13 September 2012.
25. Godoy, E.; Rosec, O.; Chonavel, T. Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 1313–1323. [CrossRef]
26. Mohammadi, S.H.; Kain, A. Transmutative voice conversion. In Proceedings of the 38th 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 6920–6924.
27. Erro, D.; Navas, E.; Hernaez, I. Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *21*, 556–566. [CrossRef]
28. Ayodeji, A.O.; Oyetunji, S.A. Voice conversion using coefficient mapping and neural network. In Proceedings of the 2016 IEEE International Conference for Students on Applied Engineering (ICSAE), Newcastle, UK, 20–21 October 2016; pp. 479–483.
29. Zhang, M.; Sisman, B.; Zhao, L.; Li, H. Deepconversion: Voice conversion with limited parallel training data. *Speech Commun.* **2020**, *122*, 31–43. [CrossRef]

30. Watanabe, T.; Murakami, T.; Namba, M.; Hoya, T.; Ishida, Y. Transformation of spectral envelope for voice conversion based on radial basis function networks. In Proceedings of the 2002 7th International Conference on Spoken Language Processing, Denver, CO, USA, 16–20 September 2002.
31. Desai, S.; Raghavendra, E.V.; Yegnanarayana, B.; Black, A.W.; Prahallad, K. Voice conversion using artificial neural networks. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taipei, Taiwan, 19–24 April 2009; pp. 3893–3896.
32. Sun, L.; Kang, S.; Li, K.; Meng, H. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 19–24 April 2015; pp. 4869–4873.
33. Wu, J.; Wu, Z.; Xie, L. On the use of i-vectors and average voice model for voice conversion without parallel data. In Proceedings of the 2016 IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Korea, 13–16 December 2016; pp. 1–6.
34. Xie, F.L.; Soong, F.K.; Li, H. A KL Divergence and DNN-Based Approach to Voice Conversion without Parallel Training Sentences. In Proceedings of the 2016 17th Annual Conference of the International Speech Communication Association (ISCA), San Francisco, CA, USA, 8–12 September 2016; pp. 287–291.
35. Sun, L.; Li, K.; Wang, H.; Kang, S.; Meng, H. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, USA, 11–15 July 2016; pp. 1–6.
36. Sun, L.; Wang, H.; Kang, S.; Li, K.; Meng, H.M. Personalized, Cross-Lingual TTS Using Phonetic Posteriorgrams. In Proceedings of the 2016 17th Annual Conference of the International Speech Communication Association (ISCA), San Francisco, CA, USA, 8–12 September 2016; pp. 322–326.
37. Hsu, C.C.; Hwang, H.T.; Wu, Y.C.; Tsao, Y.; Wang, H.M. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. *arXiv* **2017**, arXiv:1704.00849.
38. Fang, F.; Yamagishi, J.; Echizen, I.; Lorenzo-Trueba, J. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5279–5283.
39. Lorenzo-Trueba, J.; Fang, F.; Wang, X.; Echizen, I.; Yamagishi, J.; Kinnunen, T. Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data. *arXiv* **2018**, arXiv:1803.00860.
40. Kameoka, H.; Kaneko, T.; Tanaka, K.; Hojo, N. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 266–273.
41. Parsons, T. *Voice and Speech Processing*; McGraw Hill: New York, NY, USA, 1987.
42. Kominek, J.; Black, A.W. The CMU Arctic speech databases. In Proceedings of the 5th International Speech and Communication Association (ISCA) Workshop on Speech Synthesis, Pittsburgh, PA, USA, 14–16 June 2004.
43. Griffin, D.; Lim, J. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 236–243. [CrossRef]
44. Al-Dulaimi, A.W.; Moon, T.K.; Gunther, J.H. Phase Effects on Speech and Its Influence on Warped Speech. In Proceedings of the 2020 IEEE Intermountain Engineering, Technology and Computing (IETC), Orem, MD, USA, 4–5 May 2020; pp. 1–5.